Sélection automatique de variables confondantes avec l'algorithme Bayesian causal effect estimation

Denis Talbot

Département de médecine sociale et préventive, Université Laval
Unité santé des populations et pratiques optimales en santé, CHU de Québec - Université Laval

Mai 2017

Contexte

En statistique, les techniques d'**inférence causale** sont construites pour **prédire l'effet qu'aurait une intervention** potentielle, telle qu'un traitement, une campagne de santé publique ou une politique, soit à l'aide d'expériences randomisées ou de données observationnelles.

Contexte

En statistique, les techniques d'**inférence causale** sont construites pour **prédire l'effet qu'aurait une intervention** potentielle, telle qu'un traitement, une campagne de santé publique ou une politique, soit à l'aide d'expériences randomisées ou de données observationnelles.

Plus facile à réaliser à l'aide d'études randomisées...

mais parfois nécessaire d'utiliser des études d'observation.

Contexte

Pour obtenir une **estimation sans biais** de l'effet de l'exposition, il faut **ajuster pour les variables confondantes**.

Par exemple, dans un modèle de régression linéaire.

$$Y_i = \delta_0 + \beta X_i + \boldsymbol{\delta} \boldsymbol{Z_i} + \varepsilon_i$$

- Variable réponse
- Paramètre causal
- Variable d'exposition
- Variables confondantes

Contexte

Identifier les variables confondantes peut être difficile.

Une stratégie pour contourner le problème d'identification des variables confondantes $\boldsymbol{Z}$ est de mesurer et ajuster pour plusieurs variables potentiellement confondantes $\boldsymbol{U} = \{U_1, ..., U_M\}$.

Contexte

Identifier les variables confondantes peut être difficile.

Une stratégie pour contourner le problème d'identification des variables confondantes $\boldsymbol{Z}$ est de mesurer et ajuster pour plusieurs variables potentiellement confondantes $\boldsymbol{U} = \{U_1, ..., U_M\}$.

Inconvénient : L'estimateur obtenu pourrait être fortement inefficace si $\boldsymbol{U}$ est beaucoup plus grand que nécessaire.

Contexte

Identifier les variables confondantes peut être difficile.

Une stratégie pour contourner le problème d'identification des variables confondantes $\boldsymbol{Z}$ est de mesurer et ajuster pour plusieurs variables potentiellement confondantes $\boldsymbol{U} = \{U_1, ..., U_M\}$.

Inconvénient : L'estimateur obtenu pourrait être fortement inefficace si $\boldsymbol{U}$ est beaucoup plus grand que nécessaire.

Solution : Sélectionner à l'aide des données les variables confondantes

Contexte

Identifier les variables confondantes peut être difficile.

Une stratégie pour contourner le problème d'identification des variables confondantes $\boldsymbol{Z}$ est de mesurer et ajuster pour plusieurs variables potentiellement confondantes $\boldsymbol{U} = \{U_1, ..., U_M\}$.

Inconvénient : L'estimateur obtenu pourrait être fortement inefficace si $\boldsymbol{U}$ est beaucoup plus grand que nécessaire.

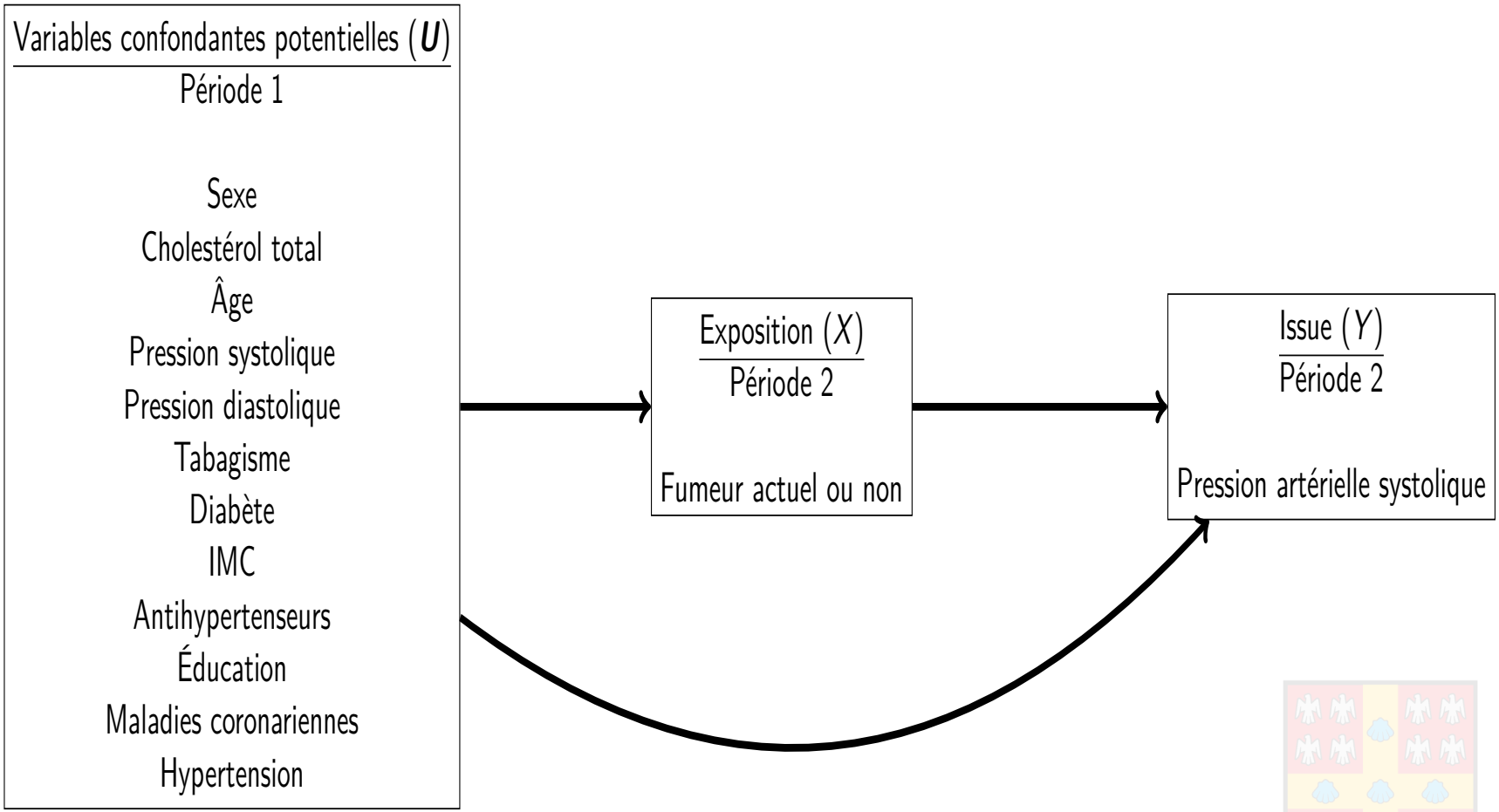
Solution : Sélectionner à l'aide des données les variables confondantes... **mais les approches classiques ne fonctionnent pas bien !**

Objectif

Illustrer l'utilisation de l'algorithme BCEE avec R pour effectuer une sélection automatique des variables potentiellement confondantes.

Le Framingham heart study (FHS)

- Étude prospective sur l'étiologie des maladies cardiovasculaires effectuée dans la communauté de Framingham, Massachuetts, aux États-Unis
- Sous-ensemble de 4 434 participants avec données anonymisées
- Deux périodes de suivi : $\approx$ 1956 et $\approx$ 1962


Variables confondantes potentielles ($\boldsymbol{U}$)
Période 1
Sexe
Cholestérol total
Âge
Pression systolique
Pression diastolique
Tabagisme
Diabète
IMC
Antihypertenseurs
Éducation
Maladies coronariennes
Hypertension
Exposition (X)
Période 2
Fumeur actuel ou non
Issue (Y)
Période 2
Pression artérielle systolique

BCEE - Intuitivement

Approche bayésienne qui permet de tenir compte de l'incertitude associée à la sélection des variables confondantes et ainsi de produire des **inférences appropriées**.

BCEE - Intuitivement

Approche bayésienne qui permet de tenir compte de l'incertitude associée à la sélection des variables confondantes et ainsi de produire des **inférences appropriées**.

Cherche à **favoriser** les modèles effectuant un ajustement suffisant pour les variables confondantes **associées simultanément à l'exposition et à l'issue.**

BCEE - Intuitivement

Approche bayésienne qui permet de tenir compte de l'incertitude associée à la sélection des variables confondantes et ainsi de produire des **inférences appropriées**.

Cherche à **favoriser** les modèles effectuant un ajustement suffisant pour les variables confondantes **associées simultanément à l'exposition et à l'issue.**

Vise également à **éviter** d'ajuster pour les variables qui ne sont qu'**uniquement associées à l'exposition**.

BCEE - Mathématiquement

Première étape - Modélisation de l'exposition

- Objectif : Identifier les déterminants potentiels de l'exposition.
- $g(E[X|\boldsymbol{U}]) = \delta_0^{\boldsymbol{\alpha^X}} + \sum_{m=1}^{M} \alpha_m^X \delta_m^{\boldsymbol{\alpha^X}} U_m$.
- Distribution *a priori* $P(\boldsymbol{\alpha^X})$ uniforme (mais autre possible).
- La distribution *a posteriori* $P(\boldsymbol{\alpha^X}|X)$ donne ces déterminants potentiels de l'exposition.

BCEE - Mathématiquement

Deuxième étape - Modélisation de l'issue

- Objectif : Estimer l'effet de l'exposition sur l'issue.
- $E[Y|X, \boldsymbol{U}] = \delta_0^{\boldsymbol{\alpha}^{\boldsymbol{Y}}} + \beta X + \sum_{m=1}^{M} \alpha_m^Y \delta_m^{\boldsymbol{\alpha}^{\boldsymbol{Y}}} U_m$.
- Distribution *a priori* $P(\boldsymbol{\alpha}^{\boldsymbol{Y}})$ informative : favorise les U_m associés à la fois avec l'exposition et l'issue ; défavorise les U_m associés uniquement à l'exposition ; non-informative pour les autres U_m.
- Inférences basées sur $P(\beta|Y)$.

Statistiques descriptives

```
CreateTableOne(vars = names(donnees)[c(4:15)],
data = donnees,
factorVars = names(donnees)[c(4, 9, 11, 12, 13, 14, 15)],
strata = "CURSMOKE2", test = F);
```

Statistiques descriptives

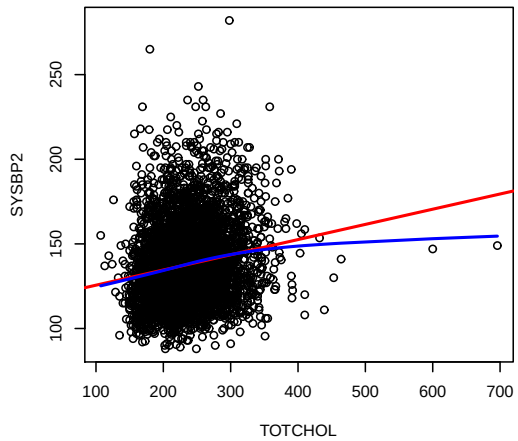
	Non-fumeurs n = 2463	Fumeurs n = 1971
Homme, n (%)	920 (37)	1024 (52)
Cholestérol, moy (é-t)	238,6 (44,4)	234,8 (44,9)
Âge, moy (é-t)	51,6 (8,7)	47,9 (8,2)
Tension systolique, moy (é-t)	135,8 (23,5)	129,3 (20,5)
Tension diastolique, moy (é-t)	84,4 (12,3)	81,4 (11,6)
Fumeur, n (%)	351 (14)	1830 (93)
IMC, moy (é-t)	26,5 (4,3)	25,0 (3,7)
Diabète, n (%)	82 (3)	39 (2)
Antihypertenseurs, n (%)	105 (4)	40 (2)
Éducation		
Secondaire terminé, n (%)	659 (27)	651 (33)
Collège non terminé, n (%)	438 (18)	301 (15)
Collège terminé, n (%)	288 (12)	224 (11)
Maladies coronariennes, n (%)	120 (5)	74 (4)
Hypertension, n (%)	917 (37)	513 (26)

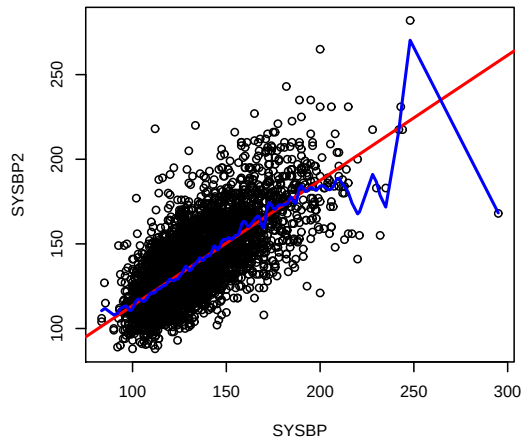
Vérifier l'adéquation des modèles

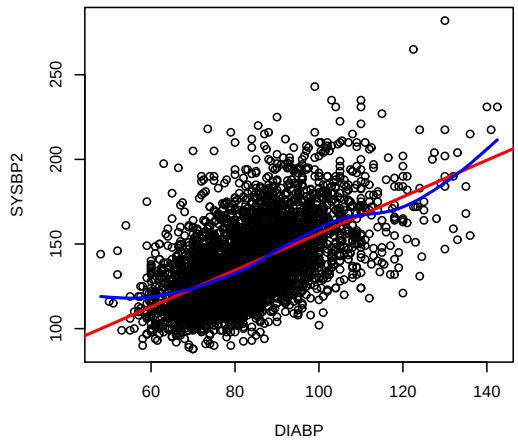
Vérifier qualitativement si la relation entre l'issue et les variables confondantes continues est linéaire.

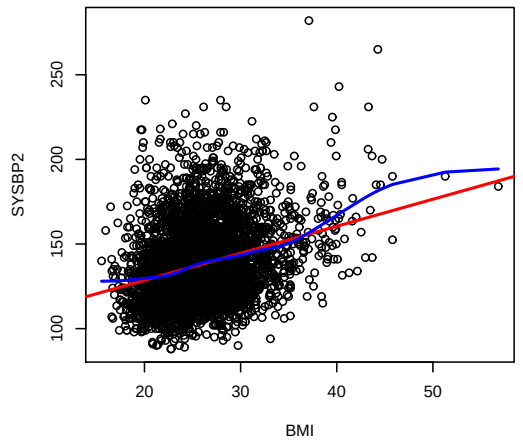
```
layout(matrix(c(1,2,3,4,5,5), ncol=2, byrow=TRUE), heights=c(4, 4, 1))

par(mai=rep(0.55, 4))
plot(donnees$TOTCHOL, donnees$SYSBP2, xlab = "TOTCHOL", ylab = "SYSBP2");
abline(reg = lm(SYSBP2~TOTCHOL, data = donnees), col = "red", lwd = 2);
lines(smooth.spline(donnees$TOTCHOL, donnees$SYSBP2), col = "blue", lwd = 2);
...

par(mai=c(0,0,0,0))
plot.new()
legend(x = "center",
 legend = c("Régression linéaire","Courbe de lissage"),
 col = c("red", "blue"), lwd = c(2,2), lty = c(1,1),
 box.lty = 0, cex = 1.5);
```

Vérifier l'adéquation des modèles









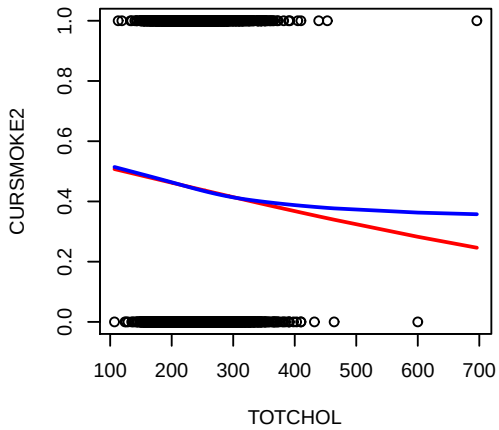
Régression linéaire
Courbe de lissage

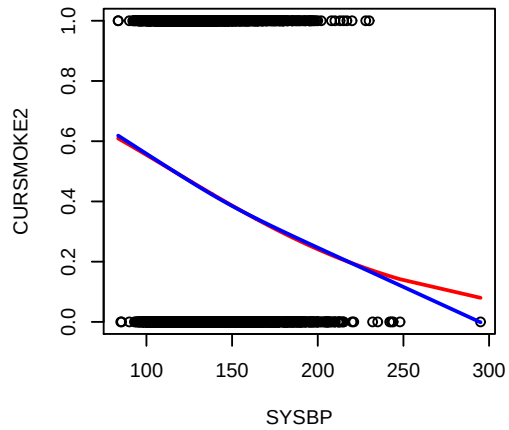
Vérifier l'adéquation des modèles

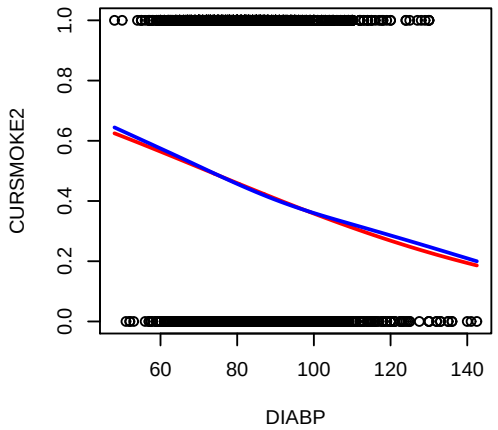
Vérifier qualitativement si la relation entre l'exposition et les variables confondantes continues est logistique.

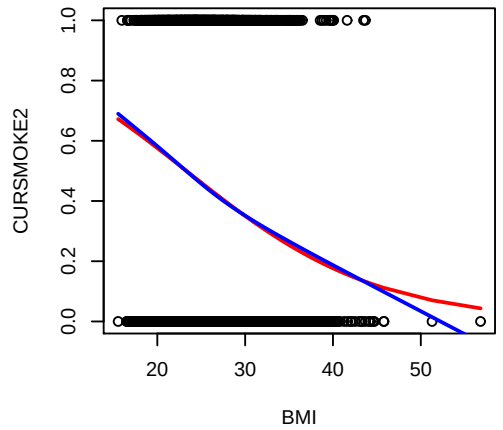
```
layout(matrix(c(1,2,3,4,5,5), ncol=2, byrow=TRUE), heights=c(4, 4, 1))

par(mai=rep(0.55, 4))
plot(donnees$TOTCHOL, donnees$CURSMOKE2, xlab = "TOTCHOL",
     ylab = "CURSMOKE2");
lines(donnees$TOTCHOL[order(donnees$TOTCHOL)],
     predict(glm(CURSMOKE2~TOTCHOL, data = donnees,
     family = binomial(link = "logit")),
     type = "response")[order(donnees$TOTCHOL)],
     col = "red", lwd = 2);
lines(smooth.spline(donnees$TOTCHOL, donnees$CURSMOKE2),
      col = "blue", lwd = 2);
```

Vérifier l'adéquation des modèles









Régression logistique
Courbe de lissage

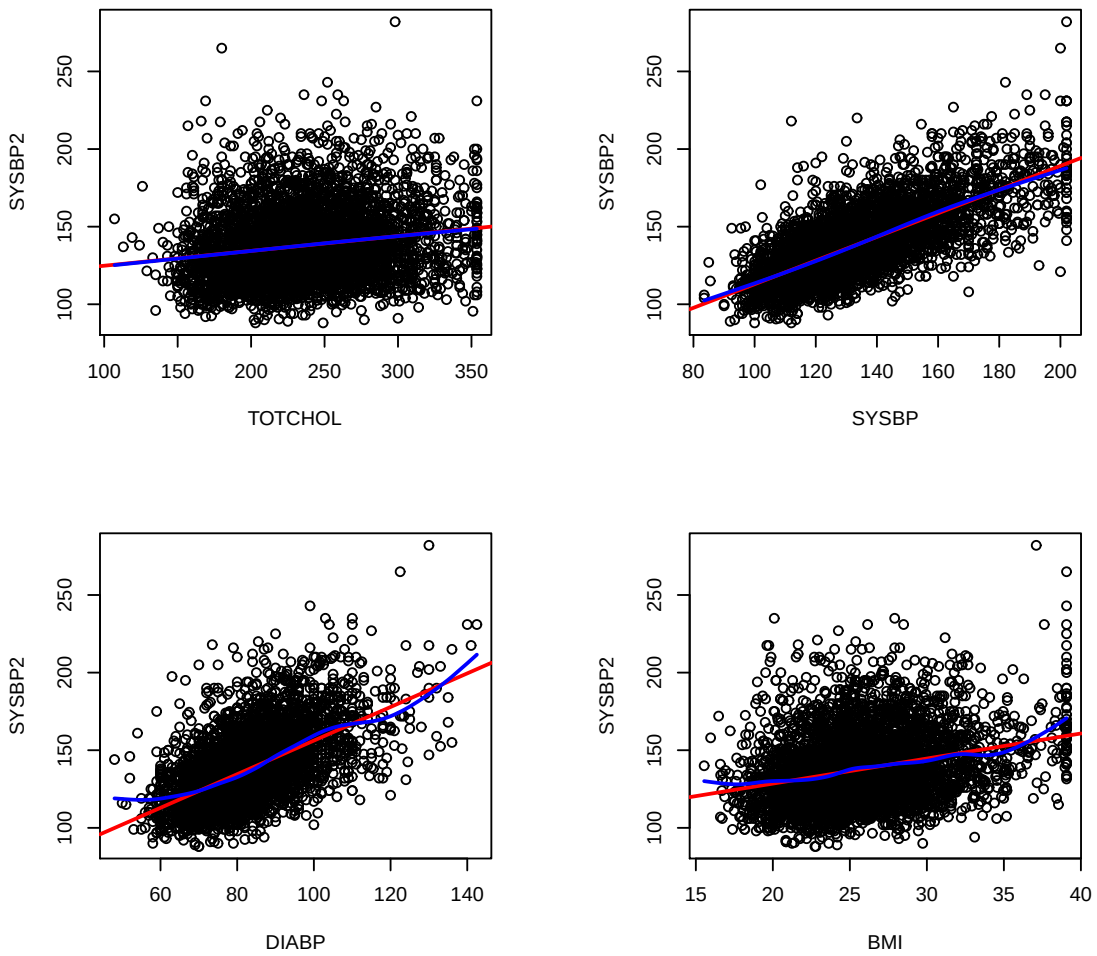
Vérifier l'adéquation des modèles

On observe des données extrêmes possiblement influentes.

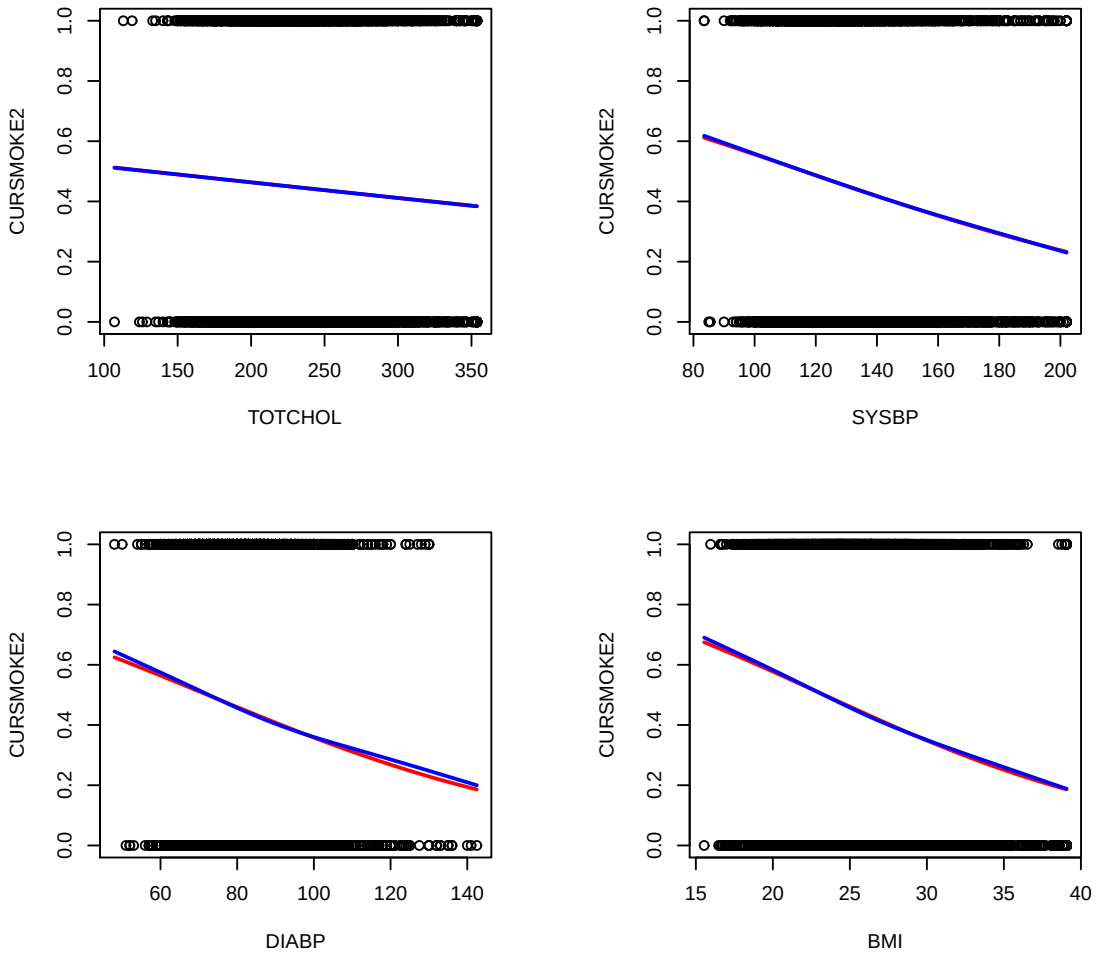
On pourrait envisager différentes solutions, j'ai opté pour tronquer au 99^e percentile.

```
donnees2 = donnees;
donnees2$TOTCHOL = pmin(quantile(donnees$TOTCHOL, 0.99), donnees$TOTCHOL);
donnees2$SYSBP = pmin(quantile(donnees$SYSBP, 0.99), donnees$SYSBP);
donnees2$BMI = pmin(quantile(donnees$BMI, 0.99), donnees$BMI);
```

Vérifier l'adéquation des modèles



Vérifier l'adéquation des modèles



Utilisation de BCEE

Les arguments essentiels de la fonction ABCEE :

- X : La variable d'exposition
- Y : La variable d'issue
- U : Les variables potentiellement confondantes, sous la forme d'une matrice
- omega : Hyperparamètre choisi par l'utilisateur
- nburn : Nombre d'itérations initiales non considérées
- niter : Nombre d'itérations post "burn-in"
- nthin : Amincissement de la chaîne de Markov
- family.X : Type de modèle pour X

Utilisation de BCEE

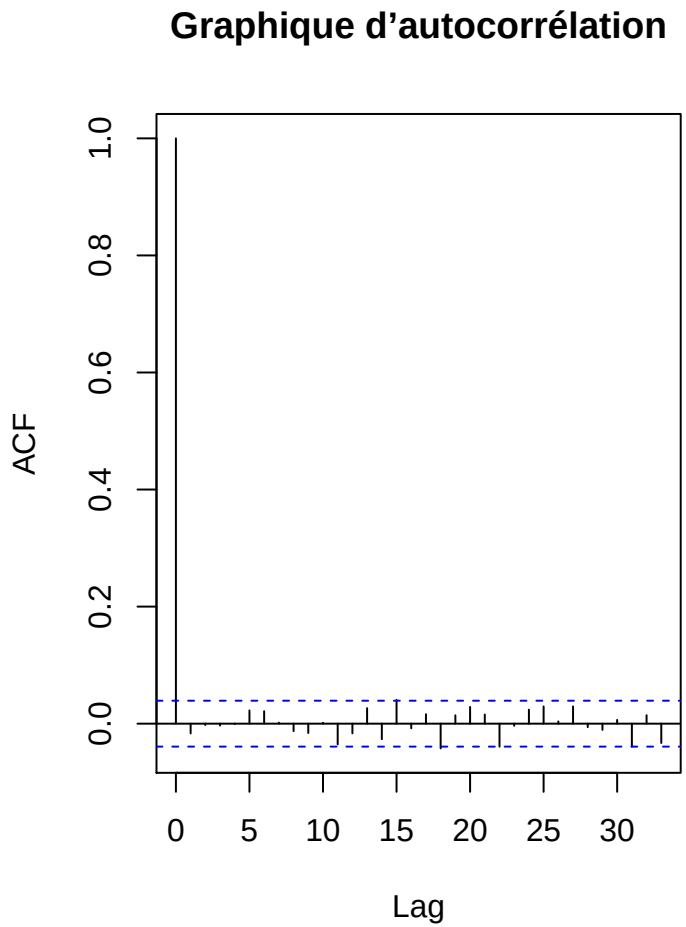
```
results = ABCEE(X = donnees2$CURSMOKE2, Y = donnees2$SYSBP2, U = U,
                omega = 500*sqrt(nrow(donnees2)), niter = 50000,
                nburn = 5000, nthin = 20,
                family.X = binomial(link = "logit"));
```

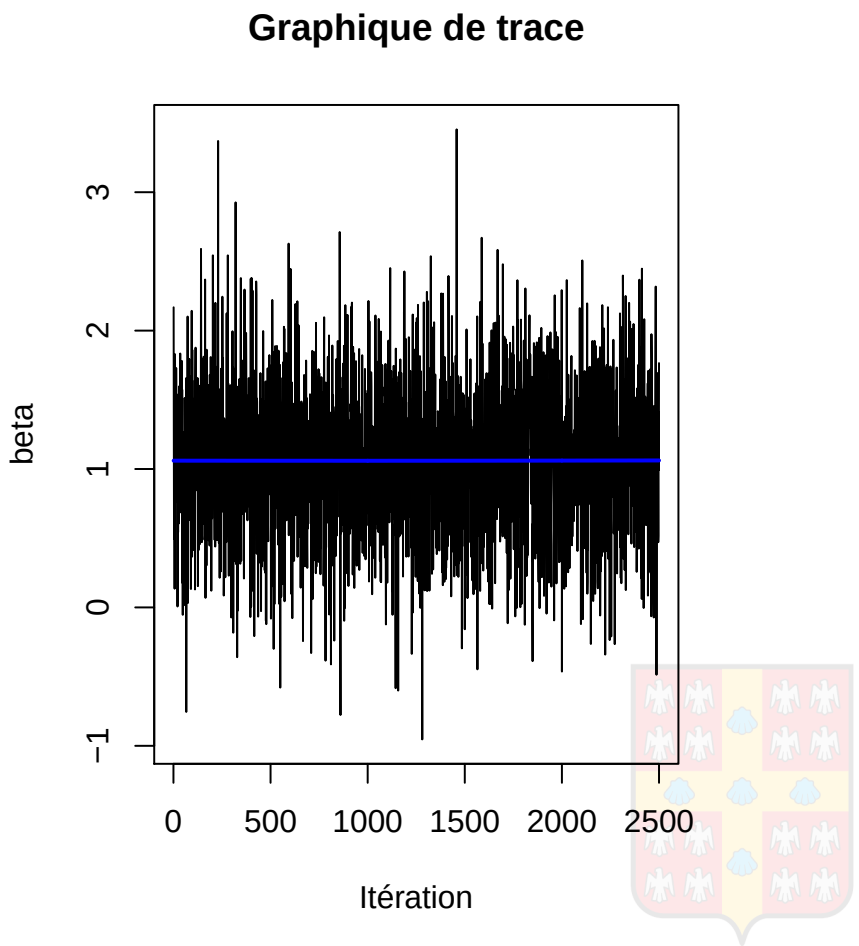
Utilisation de BCEE

Avant de considérer les résultats, il faut vérifier que les paramètres de la chaîne sont appropriés !

```
acf(results$betas, main = "Graphique d'autocorrélation");

plot(results$beta, type = "l", main = "Graphique de trace",
     xlab = "Itération", ylab = "beta");
lines(smooth.spline(1:length(results$beta), results$beta),
      col = "blue", lwd = 2);

length(results$beta);
```

Utilisation de BCEE





Utilisation de BCEE

On effectue les inférences à partir des valeurs de β échantillonnées.

```
mean(results$betas);
quantile(results$betas, c(0.025, 0.975));
```

Test-t :	-5,2 (IC à 95 % : -6,5 à -3,8 mm Hg)
Modèle complet :	1,1 (IC à 95 % : -0,4 à 2,5 mm Hg)
BCEE :	1,0 (IC à 95 % : 0,0 à 2,1 mm Hg)

Utilisation de BCEE

Quelles ont été les variables sélectionnées aux deux étapes ?

```
results$models.X;
colnames(U) = c("Sexe", "Chol", "Age", "SBP", "DBP",
                "Fum", "IMC", "Diab", "Antih", "Educ", "Coro", "Hyp");

plot(1:ncol(U), sort(colMeans(results$models.Y), decreasing = T),
 xaxt = 'n',
 xlab = "Variables potentiellement confondantes",
 ylab = "Probabilité d'inclusion",
 pch = 16);
axis(1, 1:ncol(U),
colnames(U)[order(colMeans(results$models.Y), decreasing = T)],
tick = F, cex.axis = 0.8);
abline(h = 1, lty = 2);
abline(h = 0, lty = 2);
```

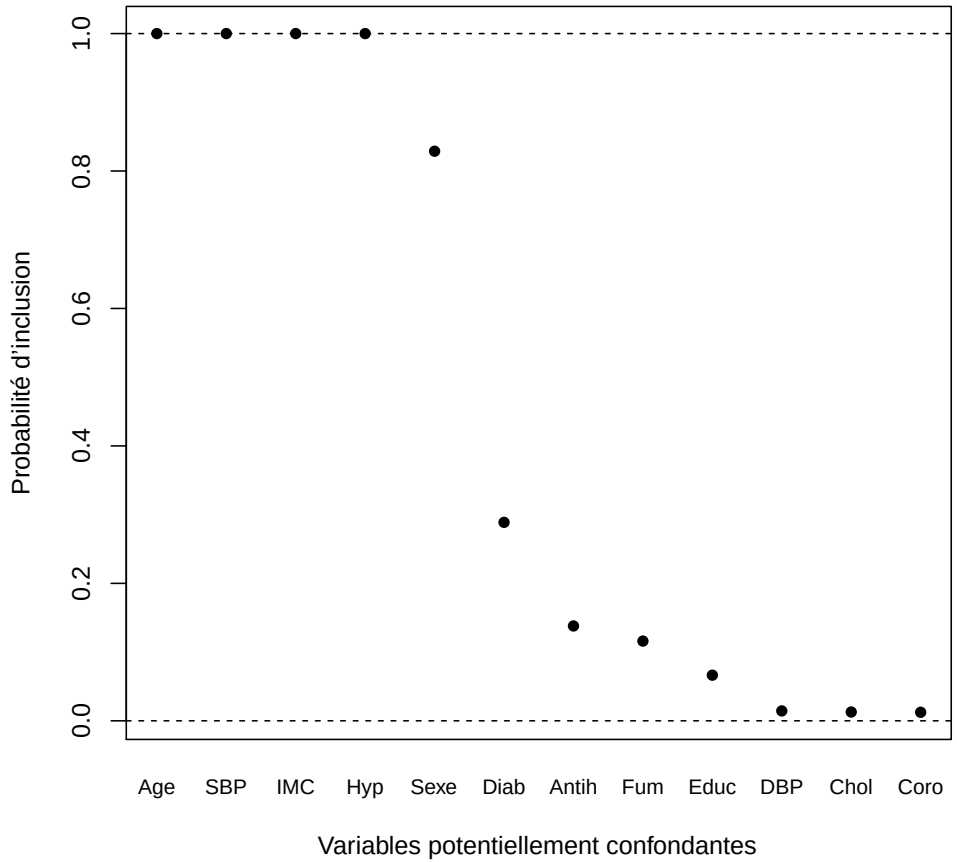
Utilisation de BCEE

Étape 1 : Déterminants de l'exposition, $P(\boldsymbol{\alpha}^{\boldsymbol{X}}|X)$

Modèle	Probabilité
Âge, Fumeur, IMC	57,4 %
Âge, Fumeur, IMC, Éducation	38,9 %
Âge, Fumeur, IMC, Pression diastolique	3,7 %

Utilisation de BCEE

Étape 2 : Variables confondantes potentielles



Conclusion

Forces

- Facile d'utilisation
- Inférences appropriées vs approches classiques
- Puissance $>$ modèle complet
- Rapide pour des utilisations usuelles

Conclusion

Forces

- Facile d'utilisation
- Inférences appropriées vs approches classiques
- Puissance $>$ modèle complet
- Rapide pour des utilisations usuelles

Limites

- Actuellement limité aux issues continues
- Actuellement limité à ≈ 20 covariables
- Pas clair comment incorporer adéquatement des termes d'ordre supérieur

Références

Talbot, D., Lefebvre, G., Atherton, J. (2015) The Bayesian causal effect estimation algorithm. Journal of Causal Inference, 3 (2), 207-236. DOI : 10.1515/jci-2014-0035

Talbot, D., Lefebvre, G., Atherton, J. (2015) The Bayesian Causal Effect Estimation Algorithm. R package version 1.1 `http://CRAN.R-project.org/package=BCEE`

Merci !

Le code `R` et la présentation seront disponibles sur
`https://sites.google.com/site/denistalbotfmed`

Détails concernant $P(\boldsymbol{\alpha}^{\boldsymbol{Y}})$:

$$P(\boldsymbol{\alpha}^{\boldsymbol{Y}}) = \sum_{\boldsymbol{\alpha}^{\boldsymbol{X}}} P\left(\boldsymbol{\alpha}^{\boldsymbol{Y}} | \boldsymbol{\alpha}^{\boldsymbol{X}}\right) P(\boldsymbol{\alpha}^{\boldsymbol{X}} | X), \text{ où}$$

$$P\left(\boldsymbol{\alpha}^{\boldsymbol{Y}} | \boldsymbol{\alpha}^{\boldsymbol{X}}\right) \propto \prod_{m=1}^{M} Q_{\boldsymbol{\alpha}^{\boldsymbol{Y}}}\left(\alpha_m^Y | \alpha_m^X\right).$$

$$Q_{\boldsymbol{\alpha}^{\boldsymbol{Y}}}\left(\alpha_m^Y = 1 | \alpha_m^X = 1\right) = \frac{\omega_m^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}}{\omega_m^{\boldsymbol{\alpha}^{\boldsymbol{Y}}} + 1}, \quad Q_{\boldsymbol{\alpha}^{\boldsymbol{Y}}}\left(\alpha_m^Y = 0 | \alpha_m^X = 1\right) = \frac{1}{\omega_m^{\boldsymbol{\alpha}^{\boldsymbol{Y}}} + 1},$$

$$Q_{\boldsymbol{\alpha}^{\boldsymbol{Y}}}\left(\alpha_m^Y = 1 | \alpha_m^X = 0\right) = \frac{1}{2}, \quad Q_{\boldsymbol{\alpha}^{\boldsymbol{Y}}}\left(\alpha_m^Y = 0 | \alpha_m^X = 0\right) = \frac{1}{2},$$

$\omega_m^{\boldsymbol{\alpha}^{\boldsymbol{Y}}} = \omega \times \left(\tilde{\delta}_m^{\boldsymbol{\alpha}^{\boldsymbol{Y}}} \frac{\sigma_{U_m}}{\sigma_Y}\right)^2$ où $\tilde{\delta}_m^{\boldsymbol{\alpha}^{\boldsymbol{Y}}}$ est le paramètre associé à U_m dans le modèle $\boldsymbol{\alpha}^{\boldsymbol{Y}}$.